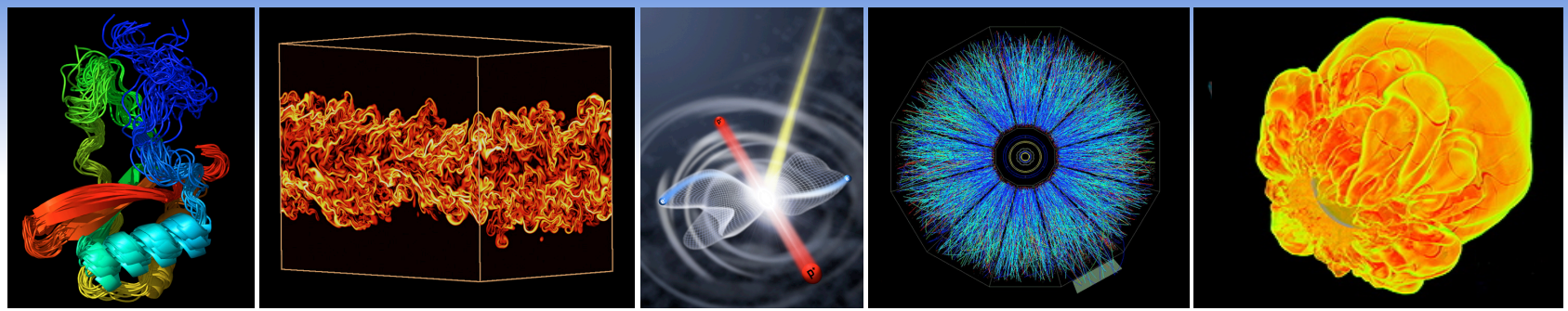


Tape Monitoring of HPSS at LBNL/ NERSC

Jason Hick

Storage Systems Group Lead

jhick@lbl.gov

February 25, 2010

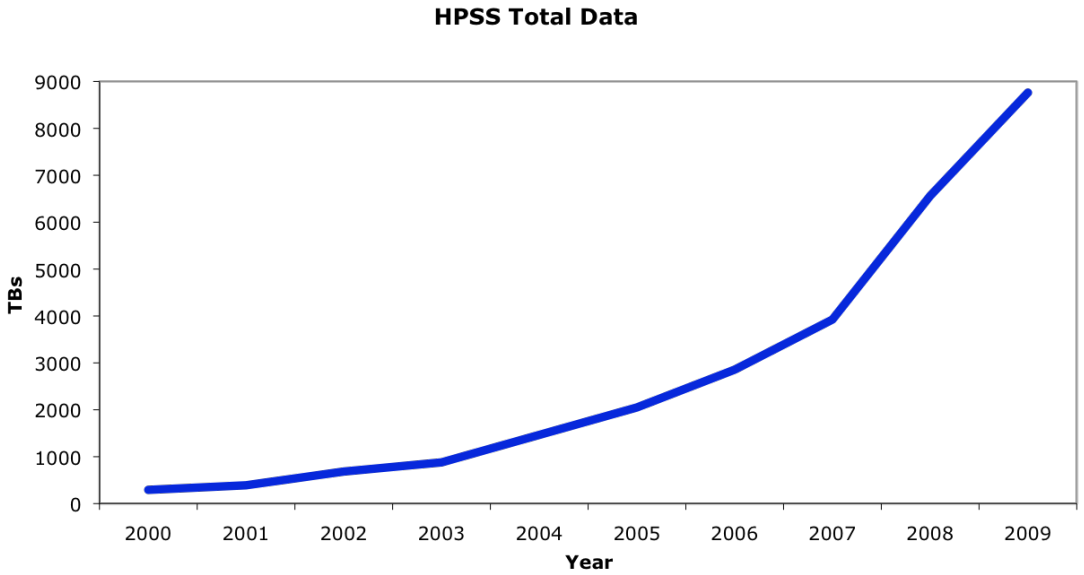
BERKELEY LAB

NERSC

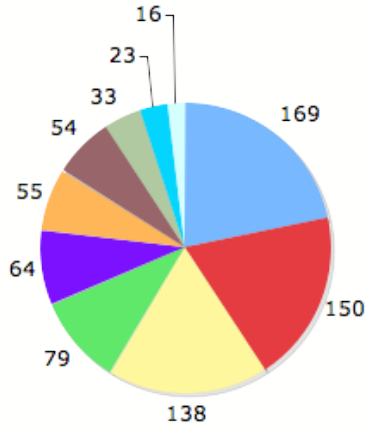
U.S. DEPARTMENT OF ENERGY | Office of Science

Tape at NERSC

- As of Feb 2010, tape holds 10 PB of data with the ability to scale to over 40 PBs
- Tape provides average compression of 40% for data stored at NERSC.
- Our average annual growth is 40-60%.
- Our media budget is approximately $500K per year.
- We use enterprise tape with a single copy of data.
- Average user file size in HPSS is 65 MB.
- 30% of IO to HPSS are reads.



2009 storage allocations categorized by area of science



Materials Science | Fusion | Chemical Science | High Energy Physics
Nuclear Physics | Climate/Environment | Computer Science | Biological Science
Math | Geo Science

U.S. DEPARTMENT OF ENERGY | Office of Science

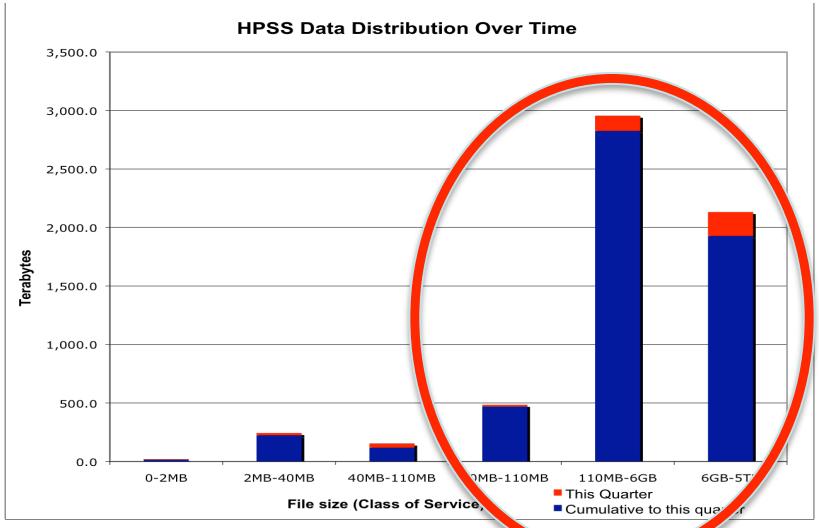
Tape Hardware & Software

- 6 x 9310 Powderhorns (read only)
  - 34 x 9840A
  - 32 x 9940B
- 4 x SL8500 (new data)
  - 84 x T10KB
  - 28 x 9840D
- Some Statistics
  - 20-40 TB I/O per day
  - 1.7 PB growth in 2009 (archive)
  - 0.5 PB growth in 2009 (backups)
- Tape related software
  - HPSS 6.2
  - ACSLS 7.3
  - Crossroads RVA/AV for tape subsystem monitoring
  - Software Delivery Platform (SDP) by Sun/STK for tape subsystem monitoring and remote resolution
  - Locally developed tape monitoring

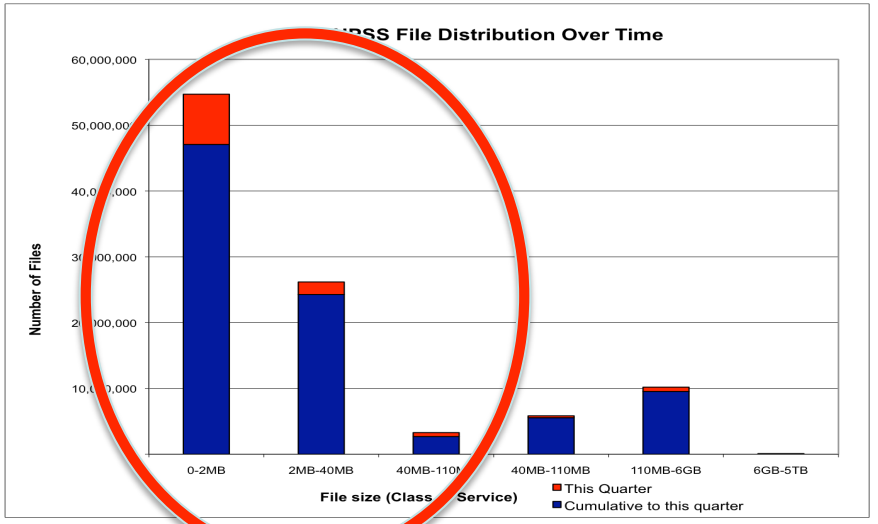
Fast Access vs. Capacity Tape

- Until another strategy proves viable (e.g. aggregation in HPSS v7), NERSC still needs both a fast access and capacity tape drive.
- 9840D fast access tape 30 seconds to first byte
- T10KB capacity tape 2 minutes to first byte

94% of data on capacity tape



83% of files on fast access tape



Our Quest in Running a Production Tape Archive

- Identify and protect against tape failures
  - Sun SDP was supposed to help with identifying problem
  - Some local solutions have helped (fault symptom code analysis, database of error reports)
- How often is hardware swapped out, and when? Do these affect error rate (i.e. if we swap out an error causing drive)?
  - Manual record keeping, helped on a few occasions, but required months to enter into a database and analyze for trends
- Is it the tape cartridge or the drive… or the combination due to variant drives?
  - A local solution (fault symptom code analysis) was most useful, but still fell short
- Match speed between disk and tape. Are we optimally configuring tape and disk resources?
  - Tape drive bandwidth determined periodically through analysis of logs and statistics
- How many tape drives by type are needed for peak ingest? (concurrent user reads/stages, migration from disk, data movement to new technology)
  - Analyze tape library manager mount logs
- Are the drives in the right location to optimize tape mount time?
  - Difficult to determine, but could analyze tape library manager mount logs
- Root cause analysis of outages (software, hardware, device, …)?
  - Manual process that took 9 months, results were mixed

Lessons Learned

- After two years of several FTEs worth of work, modest results
- Custom scripts and programs drawing on data from multiple sources and locations to maintain
- Analysis led us to make several changes in system configuration, improving user experience
- But there were many things we didn't have time for or a way to determine
  - Why is migration from our disk to tape so slow?
  - Where are the problematic drives (tape works in one drive but not another)?
  - Moving data from bad tape to good sometimes takes three or more tries before succeeding, is it the tape or the drive?

Tape Environmental Analysis

- **Provided broad set of service offerings along with system**
  - **save on precious staff time and effort**
- **Service to validate readability of the entire archive**
  - **analyzing approximately 40,000 tapes**
  - **five different generations of drives**
  - **media up to ten years old**
- **Quarterly reports to provide detailed analysis of operational performance**
  - **drives being swapped out (actual service life)**
  - **statistical determination of whether the tape or drive is problematic**
  - **tape drive bandwidth per transfer**
  - **numbers of tape drives needed for peak ingest/load**
  - **passthrough and long mount activity identified for drive relocation**
  - **preemptive media failure analysis to prioritize data movement to new media**
- **Archive requirements and usage of tape is now gaining interest in industry**
  - **systems and services are being tailored to work well for archive systems**
- **Applying the results will improve user experience with tape, improve interaction with vendor service and support, and reduce tape problems**

Quarterly Report – Example 1

Repair/Replace

- T10000B: 1,2,1,0 (572000400375)
- T10000B: 1,3,1,4 (572000400508)
- T10000B: Currently Removed (572004000429)

Watch List

- T9840D: 1,8,1,1 (5700GU004603)
- T9840D: 1,4,1,3 (5700GU003030)
- T9840D: 1,4,1,6 (5700GU003020)
- T10000B: 1,3,1,6 (572004000693)
- T10000B: 1,2,1,1 (572004000535)
- T10000B: 1,6,1,5 (572004000507)

Error Rate

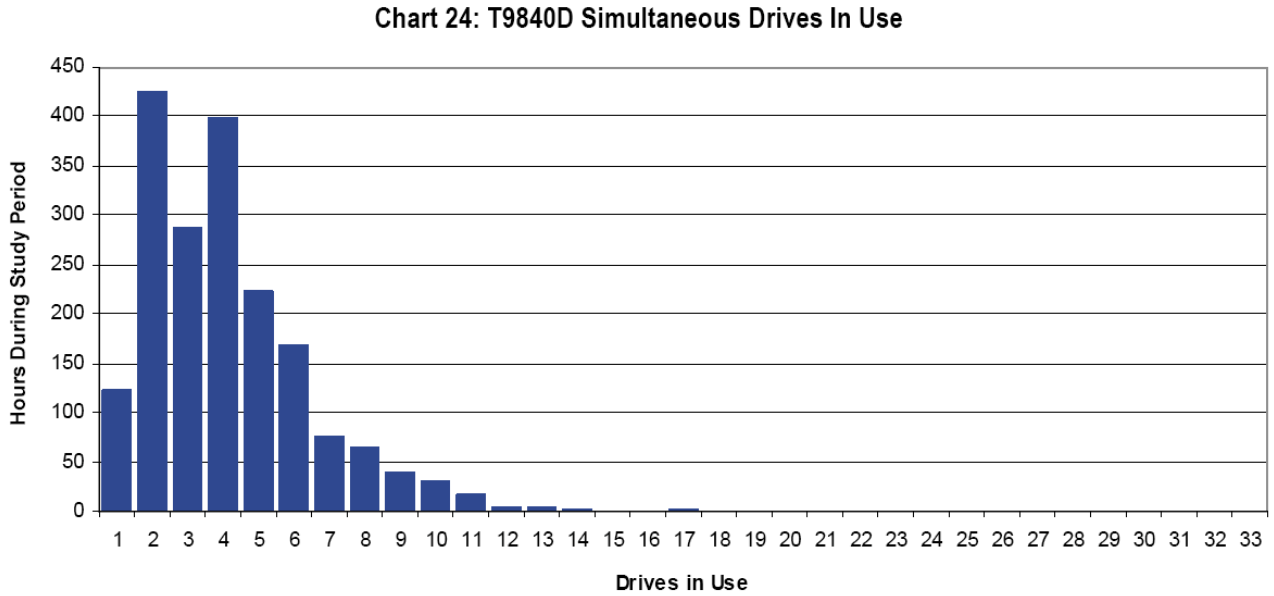
Percentage of soft errors caused by the drives on the watch and repair lists:

90%

- **Identified error producing drives**
  - **3 T10KB drives that need replacement**
  - **Addresses the most severe and important problem to us, and something**
  - **We have months of effort devoted to figuring out the same problem**
  - **Replacing should reduce soft/hard errors in next report**

Quarterly Report – Example 2

- **Identified that 9840Ds weren't being used as well as T10KBs**
  - **We identified this just prior to the report with tape type import/slot statistics that we analyze**
  - **We adjusted the size of data going to 9840D and now strike a better balance. The next report should confirm.**



Summary

- Large production tape environments are difficult to manage and scale if you don't seek answers and solutions to operational problems.
- Having an automated system to provide those answers is more effective and efficient.
- Having a detailed and well-rounded understanding of the operational tape environment leads to solutions that improve storage service to end users.